Advance in Medical Research 2024, Vol.1, No. 1

Research Paper

Machine Learning for Eye Metastasis of Primary Lung Cancer: Development and Verification of Predictive Model

Xu-Lin Liao[1]*, Sylvia Agyekum[1]*, Shi-Nan Wu[2], Qing-Jian Li[2], Jason C Yam[1#]

[1]Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, People's Republic of China

[2]Fujian Provincial Key Laboratory of Ophthalmology and Visual Science, Eye Institute of Xiamen University, School of Medicine, Xiamen University, Xiamen, Fujian 361102, People's Republic of China；

*These authors have contributed equally to this work.

#Correspondence to: Jason C Yam, **email:** soraya.jonker@mumc.nl .

Keywords: machine learning; lung cancer; extreme gradient boost; ocular metastasis; Shapley additive explanations.

Received: January 3, 2024 **Accepted:** May 14, 2024 **Published:** October 9, 2024

ABSTRACT

Background: Lung cancer (LC) is one of the fastest-growing malignancies in terms of morbidity and mortality, and ocular metastasis (OM) is a rare site of metastasis in primary lung cancer. This study aimed to construct an effective machine learning (ML)-based clinical prediction model.

Methods: We retrospectively collected clinical data from 2990 LC patients between January 2002 and December 2016 and divided them into the training and internal test sets in a 7:3 ratio. Patients were divided into the non-ocular metastasis (NOM) group and the OM group based on the presence or absence of OM. Univariate logistic regression was performed for both groups, and variables with P of < 0.05 were screened for inclusion in the ML model. We used different ML algorithms to build six ML clinical prediction models, which were internally validated by ten-fold cross-validation. The predictive performance of each model was assessed by its area under the curve (AUC), accuracy, sensitivity (recall), and specificity.

Results: A total of eight variables affecting OM in LC patients were screened by the model. The extreme gradient boost (XGB) ML model achieved optimal differential diagnostic power, and it had the best prediction performance in the internal validation set (AUC: 0.998, accuracy: 0.997, sensitivity: 0.998, specificity: 0.997). The top eight most important risk factors for OM in LC were obtained using SHAP: alpha-fetoprotein (AFP), total prostate-specific antigen (TPSA), carcinoma antigen (CA)-125, cytokeratin fragment 19 (CYFRA 21-1), CA-153, histopathological-type, CA-199, and carcinoembryonic antigen (CEA). Finally, a web-based calculator based on the XGB model was developed to predict the risk of OM in LC patients.

Conclusion: The predictive model can help identify patients with LC who are vulnerable to OM to diagnose them early and provide early personalized treatment to reduce the poor prognosis of patients developing OM and further improve their quality of life.

INTRODUCTION

According to the 2018 GLOBOCAN cancer incidence and mortality estimates published by the International Agency for Research on Cancer, lung cancer (LC) is the most commonly diagnosed cancer in both males and females (11.6% of all cases) and the leading cause of cancer deaths (18.4% of all cancer deaths)[1]. In 2020, breast cancer replaced lung cancer and became cancer with the highest accurate diagnosis rate in the world.[2] However, as of 2020, LC remains the most prevalent cancer worldwide and one of the leading causes of cancer-related deaths.[2] The main reported risk factor for LC development is still smoking. Environmental and occupational exposures, chronic lung disease, genetic susceptibility, age, gender, and race are also associated with LC development. According to the pathological type, LC can be divided into small cell carcinoma and non-small cell carcinoma (such as adenocarcinoma, large cell carcinoma, and squamous cell carcinoma). Hematoxylin-eosin staining (HE) and characteristic immunohistochemistry of lung adenocarcinoma are shown in Figure 1. These classifications are used to make treatment decisions and evaluate prognosis.[3] The two types of LC are different in sensitivity to chemotherapy and radiotherapy. At present, adenocarcinoma has surpassed squamous cell carcinoma to become the most common pathological LC type, and early metastasis is becoming increasingly common.[4]

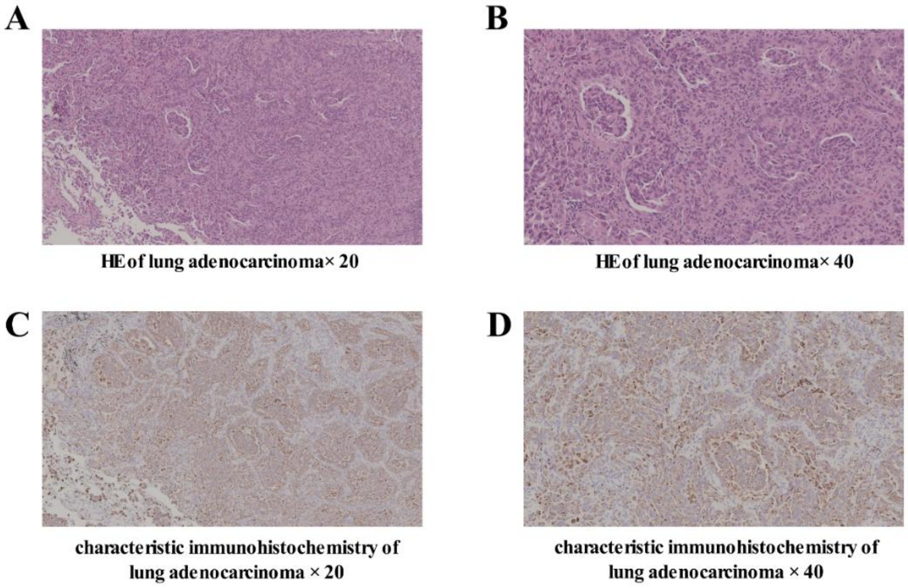

A
HE of lung adenocarcinoma× 20
B
HE of lung adenocarcinoma× 40
C
characteristic immunohistochemistry of lung adenocarcinoma × 20
D
characteristic immunohistochemistry of lung adenocarcinoma × 40

Figure 1. HE and characteristic immunohistochemistry of lung adenocarcinoma.

LC metastasis is the process by which a primary malignant tumor in the lung spreads far away from its primary site through multiple pathways. The most common sites of LC metastasis are the brain, bones, lymph nodes, and liver.[5, 6] LC metastasis is a particularly complex process that includes, but is not limited to, the involvement of LC stem cells (LCSCs) and multiple mechanisms, such as the LC microenvironment.[7, 8] Ocular metastasis (OM), as a relatively rare distant LC metastasis, tends to occur later than metastasis from other organs and tissues. Due to its low incidence and mild or even asymptomatic early clinical course, most people pay less attention to it compared to brain and liver metastases.[9] Common symptoms in patients with OM include blurred vision or even loss of vision due to macular and peripapillary retinal involvement or exudative retinal detachment. Additionally, about 12% of patients diagnosed with OM from LC will experience eye pain.[10] To date, the clinical prognosis for OM in LC patients depends mainly on the pathological tissue type, treatment modality, and the patient's preserved ocular function. Therefore, early recognition of OM for appropriate intervention is important to facilitate early disease control and improve the patient’s standard of living.

Tumor markers are certain substances produced and released by tumor cells with abnormal biochemical properties and their metabolism, often in the form of antigens, enzymes, hormones, and other metabolites within tumor cells or in host body fluids. These markers can be broadly classified into two categories: tumor cell secretion and tumor cell expression, reflecting the presence and growth of tumors. Tumor markers are mainly found in serum and cavity fluids and can be detected by various methods, such as immunology and biochemistry. Therefore, they are less difficult to collect, less expensive, and less invasive than magnetic resonance imaging (MRI) and computed tomography (CT) scans. Moreover, tumor markers can help identify or diagnose tumors just based on their biochemical or immunological properties. Recently, new tumor markers have been discovered, and their evaluation has been improved, enhancing their sensitivity and accuracy. Furthermore, tumor marker testing has become routine. However, there is currently no ideal clinical marker that can be used as an indicator of OM and LC prognosis.

Nelson et al. [11] reported a 6.7% incidence of OM in patients with LC at autopsy in 1983. Kreusel et al. found OM in approximately 7% of patients with advanced LC in 2002.[12] However, in 2008, Su et al. identified only 16 patients with symptomatic OM out of 8484 patients diagnosed with LC from 1992 to 2004.[13] In that survey, the number of OM was not significant. Alexander et al. [9] has noted that this was because the study only identified patients with symptomatic intraocular metastases and that most patients with OM were asymptomatic in the early stages. In the clinical setting, many asymptomatic patients are not easily detected. Notably, complete blindness in a specific visual field, as reported by Alexander in a 70-year-old patient with metastatic non-small cell lung cancer (NSCLC) admitted to the hospital due to progressive loss of vision after two weeks, is rarely reported. [9] This may be due to the rapid rate of progression of the patient's intraocular lesions or delayed access to the clinic.

Machine learning (ML) is a mathematical method for inductive analysis of big data and a part of artificial intelligence, which has been widely used to aid medical diagnosis and image detection.[14] Chip et al. [15]has concluded that ML classification techniques can predict the survival prognosis of LC patients, with the gradient boosting machine (GBM) model showing strong performance. Michael et al. have developed an automated deep-learning method based on chest radiograph images to identify smokers at high risk of LC.[16] Sharmila's research has demonstrated the use of ML and image processing techniques to achieve accurate LC classification and prediction.[17] However, due to the relative rarity of OM patients, few researchers have developed clinical prediction models for OM in LC patients. In this study, we wanted to identify potential biomarkers of metastatic LC by various indicators of LC patients, including general demographic data and serological indicators, and construct several clinical prediction models based on the ML approach to quantify the risk of OM based on the above-mentioned biomarkers by comparing the performance of different ML models and selecting the optimal ML model to explain the different effects of each parameter variable on LC patients. The performance of different ML models was compared, and the optimal ML model was selected to explain the different effects of each parameter variable on the occurrence of OM in LC patients. Additionally, a webpage calculator was developed to achieve personalized prediction performance for LC patients, thus, further improving the prognosis of LC patients.

MATERIALS AND METHODS

Study Population Subjects

Data for both training and internal test sets were obtained from the the Chinese University of Hong Kong. We retrospectively collected clinical data from 3019 LC patients from January 2002 to December 2016 and screened each patient for missing data. Finally, we included 2990 LC patients, including 2944 NOMs and 46 OMs, for whom information on all variables was complete. In our study, the incidence of OM was only 1.5% due to the low probability of OM in LC patients, even after 14 years of data collection on LC patients. We used a synthetic minority oversampling (synthetic minority oversampling technique SMOTE) approach to reduce the impact of unbalanced data.[18] We randomly partitioned the dataset into a training set and an internal test set in a ratio of 7:3. The clinical information exclusion criteria were 1) primary malignant or benign tumor of the eye; 2) contraindications for MRI examination; 3) other cases with unknown pathological tissue types, different tumor marker levels. The specific screening process is shown in Figure 2. This study has been approved by the Medical Ethics Committee of the The Chinese University of Hong Kong.

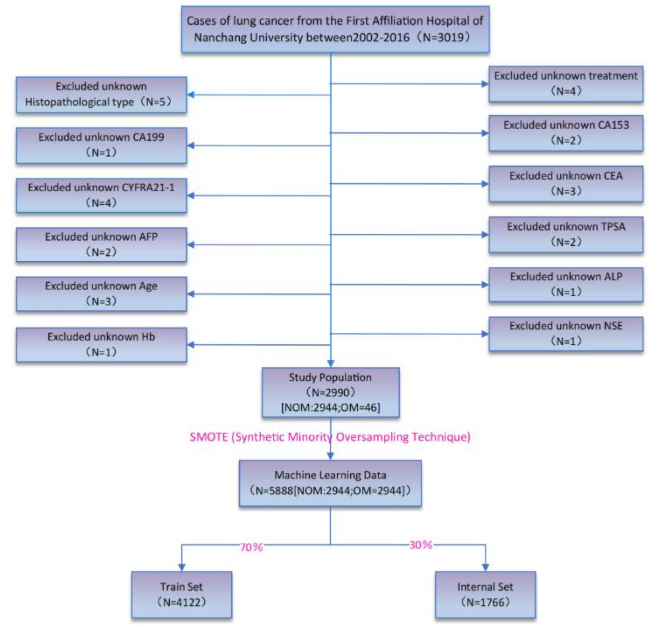

Cases of lung cancer from the First Affiliation Hospital of Nanchang University between2002-2016（N=3019）
Excluded unknown Histopathological type（N=5）
Excluded unknown CA199（N=1）
Excluded unknown CYFRA21-1（N=4）
Excluded unknown AFP（N=2）
Excluded unknown Age（N=3）
Excluded unknown Hb（N=1）
Excluded unknown treatment（N=4）
Excluded unknown CA153（N=2）
Excluded unknown CEA（N=3）
Excluded unknown TPSA（N=2）
Excluded unknown ALP（N=1）
Excluded unknown NSE（N=1）
Study Population（N=2990）[NOM:2944;OM=46]
SMOTE (Synthetic Minority Oversampling Technique)
Machine Learning Data（N=5888[NOM:2944;OM=2944]）
70%
30%
Train Set（N=4122）
Internal Set（N=1766）

Figure 2. The flow chart of data cleaning.

Data Selection

We collected various data from all clinical cases, including gender, age, and pathological tissue type, and analyzed serum tumor markers, including alkaline phosphatase (ALP), serum calcium (Ca^{2+}), hemoglobin (Hb), methemoglobin (AFP), neuron-specific enolase (NSE), carcinoembryonic antigen (CEA), total prostate-specific antigen (TPSA), carcinoembryonic antigen (CA)-125, cytokeratin fragment 19 (CYFRA 21-1), CA-153, CA-199. Univariate logistic regression analysis showed that the variables with P of < 0.05 were included in the characteristic variables of the ML model.

Statistical Analysis

We used Python (version 3.8) and R software (version 4.0.2) for the statistical analysis of the data. SMOTE technique was applied to the original dataset using Python to reduce the impact of unbalanced data on ML in dividing the dataset and subsequent validation.[17] A stratified random sampling method was used to randomly divide the SMOTE-posted dataset into the training set and the internal test set in a 7:3 ratio. We used the training set to build the model and used the internal test set to validate and evaluate the model. In this case, we de-classified the data by using the chi-square test, while the Mann–Whitney test was used for continuous, non-normally distributed data. We compared variables between patients with OM and those with NOM by using univariate logistic analysis. The eight variables with P of < 0.05 in the univariate logistic analysis were included in the construction of the ML model. Then, we performed multifactorial logistic regression (LR) to determine risk factors for OM development in LC patients. We created receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC) for subjects. The Python programming language (version 3.8) was also used to develop and evaluate ML models and design the web calculator.

Data Pre-Processing

A label-coding approach was used to address categorical variables, such as gender, treatment, and histopathological type. Univariate logistic analysis was used to select meaningful combinations of characteristics to predict the risk of developing OM in LC patients. We used the SHAP package to establish a ranking of the importance of risk factor variables for patients with intraocular metastatic LC.[19] SHAP is a method for interpreting the results of predictive models based on cooperative game theory. The method quantifies the SHAP value for each characteristic variable, representing the contribution of different characteristics to the predicted risk of intraocular metastases in LC patients. The model produces a prediction for each sample, and the sum of means of the absolute Shapley values for each feature across all samples is the overall importance score for that feature.[20] Additionally, the SHAP method demonstrates the positive or negative effect of each feature value on the predicted outcome, similar to the coefficient values in logistic regression. When the SHAP value is ＞ 0, it indicates that the corresponding feature has a higher probability of leading to a higher risk of OM, while when the SHAP value is ＜ 0, it indicates that the corresponding feature leads to a lower risk of OM.

Model building

All the algorithmic models were based on scikit-learn (Version 0.24.2). In this study, we used six different ML models: adaptive boosting (AB) model, logistic regression (LR) model, bootstrapped aggregating (BAG) model, multilayer perceptron (MLP) model, GBM model, and extreme gradient boost (XGB) model. The ML algorithm was trained and tuned to predict OM in LC patients, and the random search method in scikit-learn was used to tune the hyperparameters of the different models. Then, through the internal ten-fold cross-validation of whole data, the predictive performance of the ML models was evaluated. Finally, we selected the best-performing model by assessing AUC, accuracy, sensitivity (recall), and specificity scores to build a web calculator.

RESULTS

Demographic baseline data

The clinical data of 2990 patients were included (2202 males and 788 females), of which 46 had OM and 2944 cases had NOM. There were no statistically significant differences between the OM and NOM groups in terms of age, gender, and treatment modality (P > 0.05). In terms of histopathological subtypes, adenocarcinoma was the most common type in the two groups, and the difference in histological type distribution between the OM and NOM groups was statistically significant (P < 0.001). The differences in blood calcium, Hb, ALP, and NSE levels between the OM and NOM groups were not statistically significant (P > 0.05), while the differences in AFP, TPSA, CA-125, CYFRA 21-1, CA-153, CA-199, and CEA-related indicators were significantly different (P < 0.05). Other related indicators can be found in Table 1.

Differences in risk factors for ocular metastases

By establishing a univariate logistic regression model, we screened variables with P of < 0.05 in the univariate logistic regression analysis and multivariate logistic regression analysis to identify risk factors for the development of OM in LC patients. In the univariate logistic regression, AFP, CEA, TPSA, CA-153, CA-199, CA-125, CYFRA21_1, and pathological tissue type-related indicators were risk factors for OM (**Table 2**), and the above-mentioned indicators were included as characteristic variables in the six ML models. Multivariate logistic regression results showed that pathological tissue type, AFP, CEA, TPSA, CYFRA 21-1, and CA-153-related indicators, as risk factors for OM, were independent risk factors for the development of OM in LC patients (P < 0.05). According to the results of univariate and multivariate logistic regression, the forest maps of univariate results (figure 3A) and multivariate results (figure 3B) were drawn, respectively.

Table 1. Comparison of baseline data between the two groups

Variables	Total (n = 2990)	NOM (n = 2944)	OM (n = 46)	P value
Gender, n (%)				0.834
Female	788 (26)	777 (26)	11 (24)	
Male	2202 (74)	2167 (74)	35 (76)	
Histopathological_type, n (%)				< 0.001*
Squamous carcinoma	1139 (38)	1132 (38)	7 (15)	
Adenocarcinoma	1224 (41)	1191 (40)	33 (72)	
Large cell carcinoma	29 (1)	29 (1)	0 (0)	
Small cell lung cancer	359 (12)	355 (12)	4 (9)	
Other non-small cell lung cancer	228 (8)	227 (8)	1 (2)	
Unknown	11 (0)	10 (0)	1 (2)	
Treatment, n (%)				0.003*
untreated	164 (5)	164 (6)	0 (0)	
surgical treatment	365 (12)	365 (12)	0 (0)	
chemotherapy	1305 (44)	1279 (43)	26 (57)	
radiotherapy	46 (2)	44 (1)	2 (4)	
systemic treatment	1110 (37)	1092 (37)	18 (39)	
Age, Median (Q1,Q3)	60 (53, 68)	60 (53, 68)	59 (50, 65.75)	0.188
ALP, Median (Q1,Q3)	76 (61, 100)	76 (60, 100)	94 (66.25, 140.5)	0.005
Ca, Median (Q1,Q3)	2.25 (2.13, 2.38)	2.25 (2.13, 2.38)	2.3 (2.14, 2.43)	0.226
AFP, Median (Q1,Q3)	2.03 (0.3, 2.35)	2.01 (0.26, 2.34)	2.48 (1.97, 4.14)	< 0.001*
CEA, Median (Q1,Q3)	5.26 (2.35, 23)	5.15 (2.34, 23)	53.94 (11.7, 315.02)	< 0.001*
CA_125, Median (Q1,Q3)	22 (9, 52.18)	21.4 (9, 50.86)	118.25 (60.49, 429.15)	< 0.001*
CA_199, Median (Q1,Q3)	12 (7.74, 21.3)	12 (7.71, 21)	23.66 (10.83, 55.71)	< 0.001*
CA_153, Median (Q1,Q3)	11.12 (8.88, 21)	11 (8.66, 21)	36 (15.8, 66.26)	< 0.001*
CYFRA21_1, Median (Q1,Q3)	3.45 (2.14, 6.18)	3.45 (2.13, 5.87)	33.75 (10.76, 50.75)	< 0.001*
TPSA, Median (Q1,Q3)	1.21 (0.76, 1.67)	1.2 (0.76, 1.65)	3.52 (2.43, 5.6)	< 0.001*
NSE, Median (Q1,Q3)	16.77 (12, 23.6)	16.76 (12, 23.5)	21.9 (17, 34)	< 0.001*
Hb, Median (Q1,Q3)	120 (107, 132)	120 (107, 132)	113.5 (101, 127.5)	0.044

Notes: * p < 0.05 represented statistically significant.

Abbreviation: AFP, alphafetoprotein; TPSA, total prostate-specific antigen; CA-125, carcinoma antigen-125; CYFRA 21-1, cytokeratin fragment 19; CA-153, carcinoma antigen-153; CA-199, carcinoma antigen-199; CEA ,Carcinoembryonic antigen; Hb, hemoglobin; NSE, neuronal enolase.

Table 2. Univariate and multivariate Logistic regression

Characteristics	category	Univariate analysis		Multivarite analysis	
		OR (95% CI)	P value	OR (95% CI)	P value
Gender	Female	Ref	Ref	Ref	Ref
	Male	1.141 (0.577-2.257)	0.705	\	\
Histopathological type	Squamous carcinoma	Ref	Ref	Ref	Ref
	adenocarcinoma	4.481 (1.974-10.17)	<0.001*	2.586 (1.002-6.669)	0.049*
	large cell carcinoma	0 (0-Inf)	0.986	0 (0-Inf)	0.987
	small cell lung cancer	1.822 (0.53-6.261)	0.341	2.199 (0.595-8.121)	0.237
	other non-small cell lung cancer	0.712 (0.087-5.819)	0.752	0.269 (0.018-3.943)	0.338
AFP	\	1.441 (1.28-1.621)	<0.001*	1.411 (1.224-1.626)	<0.001 *
CEA	\	1.001 (1.001-1.002)	<0.001*	1.001 (1.001-1.002)	<0.001*
CA_125	\	1.002 (1.002-1.003)	<0.001*	1.001 (1-1.002)	0.103
CA_199	\	1.003 (1.002-1.004)	<0.001*	1.001 (1-1.003)	0.105
CA_153	\	1.014 (1.01-1.018)	<0.001*	1.011 (1.006-1.015)	<0.001*
CYFRA21_1	\	1.008 (1.005-1.012)	<0.001*	1.006 (1.001-1.011)	0.029*
TPSA	\	1.699 (1.522-1.898)	<0.001*	1.726 (1.52-1.96*1)	<0.001*
Gender	\	1.141 (0.577-2.257)	0.705	NA	NA
Age	\	0.984 (0.958-1.01)	0.235	NA	NA
ALP	\	1.003 (1-1.005)	0.064	NA	NA
Ca	\	2.137 (0.689-6.63)	0.188	NA	NA
Hb	\	0.986 (0.971-1)	0.057	NA	NA
NSE	\	1.002 (0.996-1.008)	0.507	NA	NA

Notes: * p < 0.05 represented statistically significant.

Abbreviation: AFP, alphafetoprotein; TPSA, total prostate-specific antigen; CA-125, carcinoma antigen-125; CYFRA 21-1, cytokeratin fragment 19; CA-153, carcinoma antigen-153; CA-199, carcinoma antigen-199; CEA ,Carcinoembryonic antigen; Hb, hemoglobin; NSE, neuronal enolase.

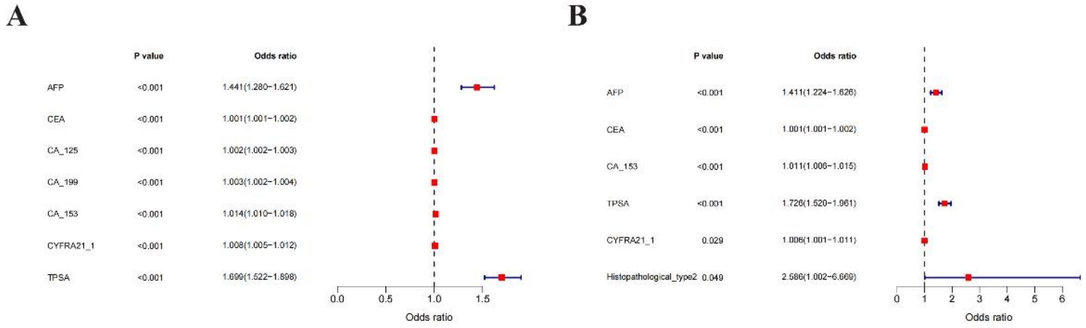


Figure 3. Univariate and multivariate logistic regression results forest map.

(A) The univariate logistic regression results forest map. (B) The multivariate logistic regression results forest map.

Model performance

We assessed the risk probability of developing OM in LC patients by six different ML models built for MLP, AB, BAG, LR, GBM, and XGB and the associated accuracy comparison. The prediction performance of all models was evaluated with 10-fold cross-validation of the whole data set, as detailed in Figure 4A.

The XGB model was shown to perform best in the validation set results with an AUC of 0.998, an accuracy of 0.997, a sensitivity of 0.998, and a specificity of 0.997 (Table 3). The results of the 10-fold cross-validation showed that XGB had an AUC of 0.999 and a standard error of 0.001, outperforming other ML models and traditional LR. The model validation results based on the validation set are shown in Figure 4B, where the XGB model continued to perform best in predicting the occurrence of OM in LC (AUC = 0.998).

We constructed an ROC curve based on the optimal ML model XGB to assess the stability of its results, as shown in Figure 4C. Five-fold cross-validation was performed to assess the stability and accuracy of the XGB, and our results showed that the XGB had good stability. Moreover, the confusion matrix for each ML model result was plotted against the SMOTE balance data, as shown in Figure 5. The number of accurately predicted OM samples in the XGB ML algorithm was 2944 cases, and the number of accurately predicted NOM samples was 2938 cases. For the above-mentioned six ML models, we plotted the maximum values of the five metrics evaluated by the radar plot, where XGB had the best value in each metric evaluation compared to other models for sensitivity, F1 score, AUC, accuracy, and specificity (Figure 6**).**

Table 3. Comparison of six machine learning metrics

Model	F1	AUC	Accuracy	Sensitivity	Specificity
AB	0.943	0.989	0.943	0.942	0.944
LR	0.908	0.954	0.908	0.899	0.918
BAG	0.952	0.99	0.952	0.967	0.937
MLP	0.927	0.962	0.927	0.931	0.922
GBM	0.988	0.996	0.988	0.997	0.978
XGB	0.999	0.998	0.997	0.998	0.997

Abbreviation: AUC, area under the curve; AB, adaptive boosting; LR, logistic regression; BAG, bootstrapped aggregating; MLP, multilayer perceptron; GBM, gradient boosting machine; XGB, extreme gradient boost.

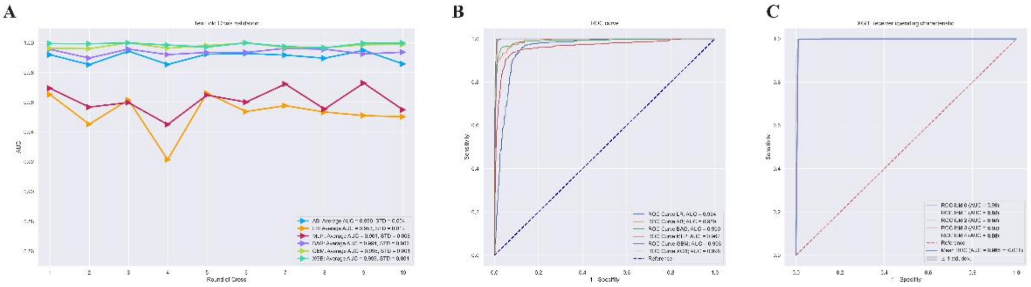


Figure 4. Validation of machine learning algorithms.

(A) AUC values of 10-fold cross-validation. (B) validation of machine learning algorithms. (C) ROC curve in the XGB model. (D) AUC is used as an indicator of performance, and the XGB model achieved the best predictive performance.

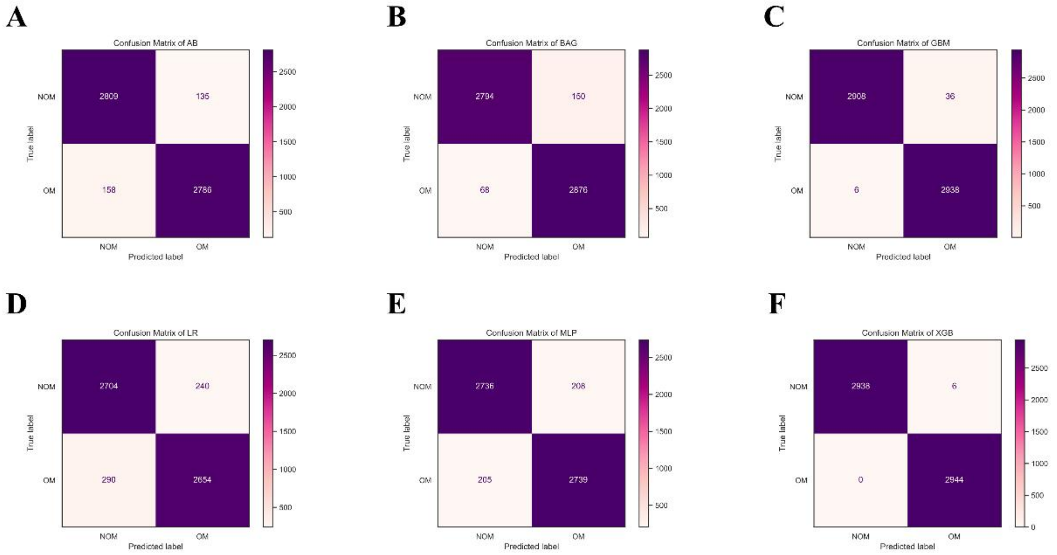


Figure 5. Confusion matrix of six machine learning models.

(A-F) The best correct classification (accuracy) of OM for the machine model was XGB.

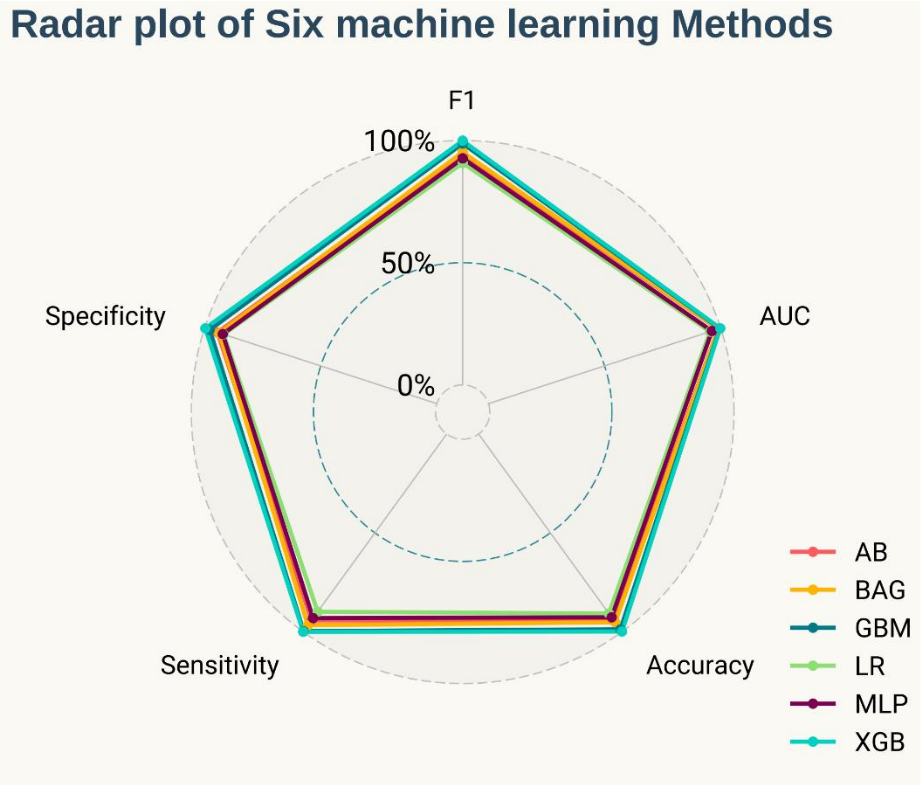


Figure 6. Radar plot of six machine learning methods.

Among the six machine learning models, XGB showed the best performance in F1 score, AUC, accuracy, sensitivity, and specificity.

Importance of characteristic variables

We ranked the importance of risk factors for whole subjects by SHAP. Additionally, we took one low-risk subject and one high-risk subject and analyzed two typical cases to demonstrate the interpretability of the models.

We used the SHAP library to build a risk factor model for OM in LC patients based on the XGB (Figure 7A). According to Figure 7, the global interpretation of the risk factor model obtained by SHAP was as follows. Among the SHAP values on the X-axis in Figure 7, all values on the left side are the proportion of predicted values that are negatively correlated, and the values on the right side are the proportion of predicted values that are positively correlated. The Y-axis represents the descending order of importance of the effect these characteristics have on OM in LC patients. In the model for XGB, the variables are, in order of importance, AFP, TPSA, CA-125, CYFRA 21-1, CA-153, histopathological type, CA-199, and CEA, with their details demonstrated in Figure 7B. Furthermore, according to the SHAP library, we sampled two subjects each, which included members of the OM and NOM groups. The base value calculated from our model was −13.53, where the output value of the low-risk group was −22.9, and that of the high-risk group was −9.06 **(**Figure 7C, D**)**. For the low-risk group, AFP, TPSA, CA-125, and CEA were low-risk factors for OM, while other variables, such as CYFRA 21-1 and CA-199, were high-risk factors. For the high-risk group, AFP, histopathological type, CA-125, CA-153, and CYFRA21-1 were high-risk factors, while TPSA and CA-199 were low-risk factors. In both subject samples, CYFRA21-1 was a high-risk factor, and TPSA was a low-risk factor for OM occurrence. Other details of specific values are shown in Figure 7.

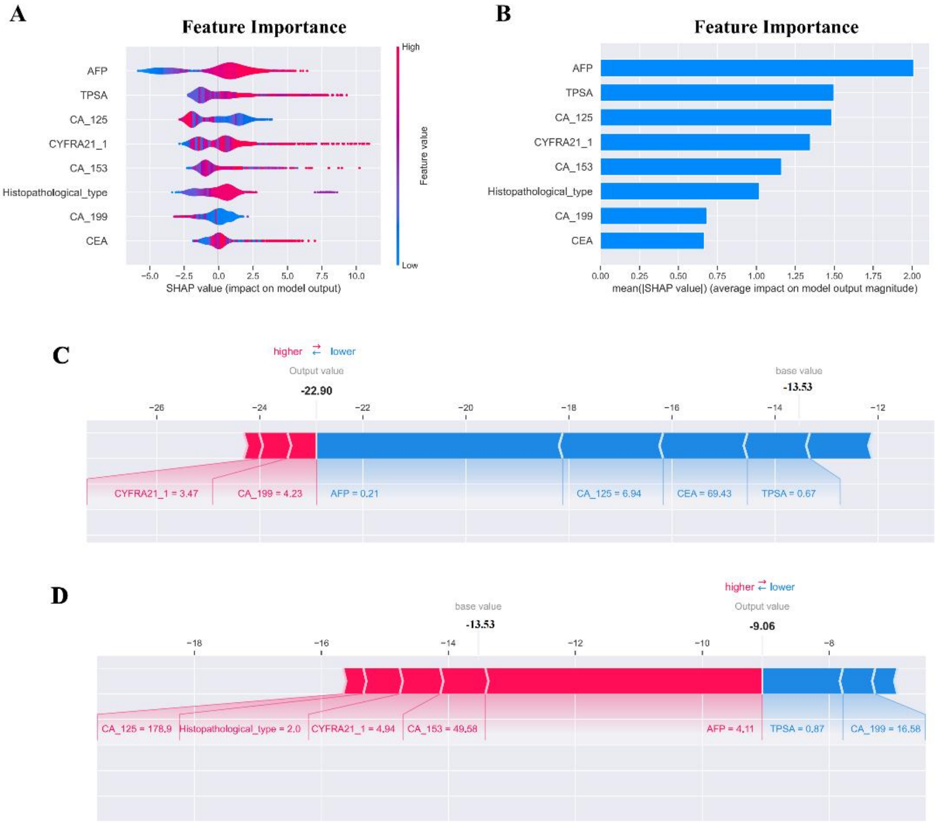
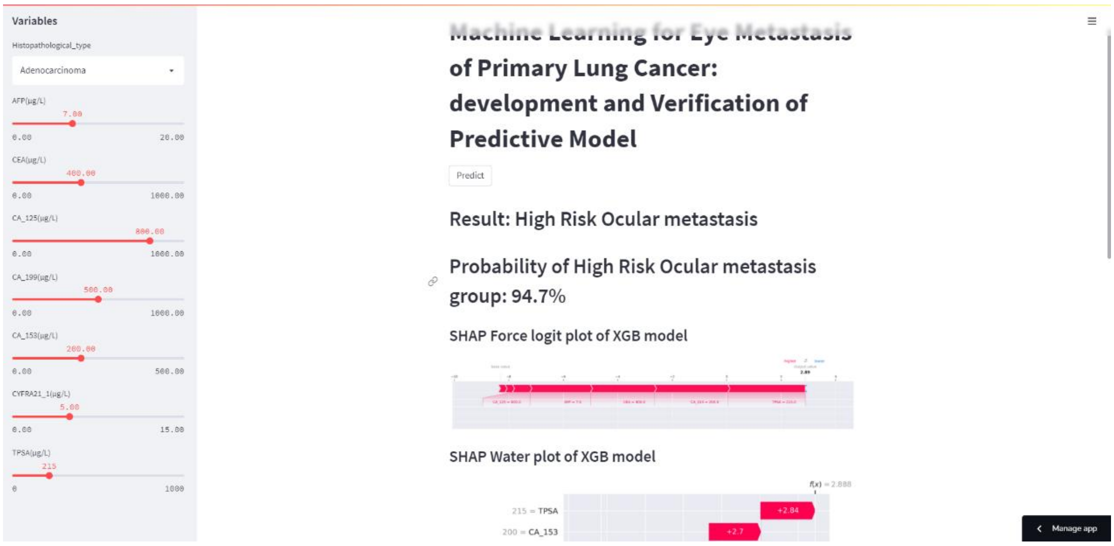


Figure 7. SHAP summary plot and SHAP model explanation of two typical predictions.

(A) The features are ranked according to the sum of the SHAP values of all patients, and the SHAP values are used to show the distribution of the influence of each feature on the output of the XGB model. The horizontal axis represents the SHAP value corresponding to the feature, with a positive SHAP value helping to predict OM. (B) Descending histogram of mean importance values calculated according to characteristic variables. (C) the low-risk SHAP interpretation model of eye metastasis in lung cancer patients. (D) the high-risk SHAP interpretation model.

Web page calculator

The XGB model had optimal prediction performance; hence, the above web predictor can be used to predict the risk probability of OM. Users only need to enter the specific number of characteristic variables in the sidebar of the web page and click predict to obtain the risk probability of OM. The risk value can be calculated in real time according to the input variables of the user, and the degree of influence of each variable on the risk of eye metastasis can be sorted (https://shimunana-lungcancer-eye-lung-cancer-eye-1iank4.streamlit.app/ (Figure 8).



Variables

Histopathological_type: Adenocarcinoma

AFP(μg/L): 7.00 (0.00–20.00)

CEA(μg/L): 400.00 (0.00–1000.00)

CA_125(μg/L): 800.00 (0.00–1000.00)

CA_199(μg/L): 500.00 (0.00–1000.00)

CA_153(μg/L): 200.00 (0.00–500.00)

CYFRA21_1(μg/L): 5.00 (0.00–15.00)

TPSA(μg/L): 215 (0–1000)

Machine Learning for Eye Metastasis of Primary Lung Cancer: development and Verification of Predictive Model

Result: High Risk Ocular metastasis

Probability of High Risk Ocular metastasis group: 94.7%

SHAP Force logit plot of XGB model

SHAP Water plot of XGB model

Figure 8. Web calculator for predicting ocular metastasis of lung cancer based on extreme gradient boosting model.

The URL is https://shimunana-lungcancer-eye-lung-cancer-eye-1iank4.streamlit.app/.

DISCUSSION

In this study, six ML algorithms were used for the first time to predict the OM in LC patients, and an XGB model, which can be used to clinically predict OM occurrence in LC, was obtained and interpreted. Compared with the conventional LR model, the XGB with the base learner as a tree model had higher prediction accuracy, and the quantitative analysis of the feature contribution effectively assisted the process of feature selection. Subsequently, we designed a network risk calculator based on the XGB model to estimate the probabilistic risk of OM in LC patients to help clinicians develop targeted diagnostic and treatment plans, making precision treatment possible.

LC is the malignant tumor that has the highest incidence and mortality rate in the Chinese region and often metastasizes distantly. Thus, many LC patients in China have already developed local or distant metastases by the time they are diagnosed. Although OM is relatively rare, it predicts a poor prognosis for LC patients. Therefore, early diagnosis of OM is essential. Tumor marker evaluation is currently of great practical value in aiding the diagnosis of tumors, analyzing the course of the disease, guiding treatment, monitoring compound, and judging prognosis, as well as evaluating the efficacy of treatment and following up and observing high-risk groups. When tumor markers reach a certain level, the possibility of tumor metastasis can be predicted, which compensates for the limitations of CT, positron emission tomography (PET)/CT, and MRI. In our analysis of clinical data from 2990 LC patients, binary logistic regression analysis showed that AFP, TPSA, CA-125, CYFRA 21-1, CA-153, pathological tissue type, CA-199, and CEA might be independent risk factors for OM in patients with multicellular LC ($P < 0.001$). For the possibility of OM in LC, the serum levels of TPSA, AFP, CYFRA 21-1, CA-125, and CA-153 showed greater importance.

AFP is a glycoprotein derived from embryonic endodermal tissue cells and is present in higher levels in fetal serum than in adults. Cellular immunity is the main immune mechanism against cancer, and natural killer cells, dendritic cells, and T lymphocytes are involved in immune surveillance.[21] AFP has been shown to influence three important immune cells to exert anti-tumor effects.[22, 23] AFP is often used as a diagnostic marker for hepatocellular carcinoma, and its level can become abnormal 6–12 months earlier compared to signs of cancer on imaging, providing an important basis for early diagnosis of hepatocellular carcinoma. AFP, which is modified in vitro to enhance immunogenicity and immune response, has become a target for immunotherapy of hepatocellular carcinoma.[24] In 2021, Jing Tang et al. found that AFP was an independent risk factor for LC metastasis and showed the highest sensitivity relative to other tumor markers.[25]

TPSA is considered an effective means of detecting and predicting prostate cancer. In 2021, Čamdžić et al. analyzed prostate puncture specimens from 115 patients with normal pre-treatment TPSA levels and showed that TPSA values were positively associated with the prevalence of prostate cancer found in puncture biopsy specimens and that elevated TPSA values increased the likelihood of prostate cancer.[26] Ge et al. have identified serum concentrations of TPSA as an independent risk factor for OM in elderly LC patients and that the combination of CA-125, CA-153, and TPSA was an accurate predictor of OM in elderly LC patients.[27]

CA-125, also known as mucin 16, is a membrane glycoprotein belonging to the extensive mucin family and is encoded by the gene with the same name, MUC16. CA-125 has been used as a tumor marker for ovarian cancer following its discovery in certain patients with specific cancer or other benign diseases.[28] In 2020, Wang et al. found that the combination of NSE and CA-125 could aid in the prediction of liver metastases in LC, providing improved diagnostic accuracy.[29] In 2021, Manoj K Bind et al. demonstrated that CA-125 could be a good adjunct to diagnose cases of gallbladder cancer as well as to imaging studies.[30] However, due to its limited specificity and sensitivity, CA-125 alone is still not an ideal biomarker. In recent years, clinical practitioners have used CA-125 in combination with HE4, another marker for ovarian cancer that has recently been introduced into clinical use, to improve clinical performance. In this way, better sensitivity and specificity can be achieved in identifying recurrences of epithelial ovarian cancer.[28] Additionally, better sensitivity and specificity in identifying the recurrence of epithelial ovarian cancer are possible.

CA-153 is secreted by epithelial cells with secretory function, such as breast, lung, and intestinal cells, and can also be detected in normal human excreta. It has been first found in the membrane of breast cancer cells, consisting of three structures: the membrane region, the intracellular region, and the extracellular region rich in glycosyl groups.[31] CA-153 can be separated from the cancer cell membrane and released into the bloodstream, and its sensitivity in advanced breast cancer can reach 80%. Furthermore, CA-153 has a certain positive rate in other malignant tumors, such as LC, colon cancer, pancreatic cancer In contrast, the antigenic determinant cluster of its extracellular region can be determined by specific binding to monoclonal antibodies. Additionally, the CA-153 level is abnormal in lung, endometrial, and gastrointestinal cancers. Biao et al. suggested that smoking can be associated with LC and eye lesions by altering CA-153 as a risk factor. When serum levels of CA-153 are more than 22.33 U/ml, CT or MRI should be performed to detect OM.[32]

CYFRA21-1 is a soluble fragment of cytokeratin 19 produced by cancer cells during the differentiation process and is mainly found in the cytoplasm of compound tumor epithelium. It has greater significance for LC diagnosis.[33] When cells die, CYFRA 21-1 is released into the bloodstream as a cleaved fragment, resulting in elevated CYFRA 21-1 serum levels. According to previous studies, CYFRA 21-1 can be used as a valid indicator for the diagnosis of bladder cancer.[34] It has also been associated with gastrointestinal and gynecological tumors such as epithelial ovarian cancer.[35] Hiromichi et al. reported that high levels of CYFRA 21-1 were associated with advanced stages of LC tumors.[36] CYFRA21-1 was abundant in LC tissues, especially in lung squamous carcinoma, where it was highly expressed. Thomas et al. found that CYFRA 21-1 could be used in the diagnosis, prognosis, and monitoring of non-small cell LC (NSCLC).[37] Jing Tang et al. concluded from their analysis that CYRFA21-1 is an independent risk factor for LC metastasis and that CYRFA21-1 has the highest area under the ROC curve values and better sensitivity and specificity values, suggesting that CYFRA21-1 has better diagnostic value compared to other tumor markers.[25] In 2019, Qi Lin et al. showed that the combination of CYFRA21-1 and CA-153 had high accuracy, sensitivity, and specificity in predicting OM.[38] Recent studies have shown that age and smoking status can influence serum cytokeratin 19 fragment levels in cancer-free individuals and that high levels of serum CYFRA 21-1 are associated with older age and smoking.[39]

ML is a mathematical model that applies artificial intelligence in the context of big data to obtain the relationship between variables from a large number of data samples. ML has been closely integrated with medicine in recent years and has gradually given rise to a medical-industrial crossover direction, which is one of the important branches of data mining. Up to now, few ML models have been applied in the direction of OM in LC. In our study, we compared six different ML models to predict the risk probability of OM in LC patients, comparing the F1 score, sensitivity, specificity, AUC, accuracy, and other indicators. Finally, the XGB model could achieve optimal performance. The XGB model is an improvement of the traditional ML algorithm gradient boosting decision tree (GBDT).[40] The basic idea is that in each new cycle of computation, the residuals from the previous cycle are used as new data for learning, and a weak classifier is generated in the negative gradient direction to minimize the residuals of the current cycle. These weak classifiers are then accumulated to generate a strong learner to fit the global truth values. Compared to GBDT, XGB introduces a regular term in the objective function to prevent overfitting of the model during training and improve the robustness of the model.

Although ML models are more powerful and relatively more accurate than traditional statistical models, the interpretation of the models is correspondingly more complex, as are black boxes, limiting their further clinical application. In our study, we interpreted the optimal ML model XGB utilizing SHAP, a stand-alone ML model interpretation technique that can interpret both global and individual sample black box models and help understand the relationship between predictors and outcomes in MLP models. Therefore, our study, based on the selection of the optimal ML model, aimed to enhance the global interpretation of XGB applied to the prediction of the risk of developing ocular metastases in LC patients, which would help improve clinicians' confidence in the clinical application of our ML model, assist clinicians in providing personalized treatment plans during the consultation and treatment of patients, and provide technical support for clinical decision-making.

What remains to be improved is that there were still some limitations in our current study. First, the sample size of the OM group was relatively small, and all participants were from the same region and hospital; thus, the results were not sufficiently convincing. This was a single-center retrospective study, and the performance of the ML algorithm might vary according to the characteristics of patients in different regions and the data sets of different institutions. Therefore, in our next study, we will try to obtain a large multi-center sample dataset to validate the robustness and reproducibility of our model. Second, with the relatively small number of variables incorporated for learning from the characteristic variables in our ML algorithm, we will incorporate as many clinical indicators as possible in our subsequent studies and perform prospective validation based on this model with larger sample sizes to continue exploring the key risk factors for the development of OM in LC patients and further modify various parameters of the model to improve the accuracy of the XGB prediction model.

CONCLUSION

The current study included data on the characteristics of 2990 LC patients and built a prediction model for the risk of developing OM in LC patients according to the XGB model, demonstrating that the XGB model performed best among the six ML models. The prediction model can help identify LC patients at high risk of developing OM, provide early and personalized diagnosis and treatment plans, and assist clinicians with technical support when making clinical decisions for patients, thereby reducing the serious consequences of OM in LC, further improving patients' prognosis and quality of life, coordinating the rational use of healthcare resources, and reducing the burden on society.

ACKNOWLEDGEMENTS

Not applicable.

FOUNDING

National Natural Science Foundation (No: 82160195); Jiangxi Province Double Thousand Plan Science and Technology Innovation High-end Talent Project (2022); Major (Key) R&D Program of Jiangxi Province (No: 2022103; 20181 bbg70004; 20203BBG73059); Excellent Talents Development Project of Jiangxi Province (No.: 20192BCBL23020)

AVAILABILITY OF DATA AND MATERIALS

The datasets used and/or analyzed during the present study are available from the corresponding author on reasonable request.

COMPETING INTERESTS

This study did not receive any industrial support. The authors have no competing interests to declare regarding this study.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by XL. Dr Jason CY is the guarantor of integrity of the entire study. The first draft of the manuscript was written by XL. All authors commented on previous versions of the manuscript. The statistical analysis was performed by XL. Clinical data was collected by SA, Literature research was performed by SW and QL. All authors read and approved the final manuscript.

REFERENCE

1. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. 68(6): p. 394-424.
2. Cao, W., et al., *Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020.* Chin Med J (Engl), 2021. 134(7): p. 783-791.
3. Collins, L.G., et al., *Lung cancer: diagnosis and management.* Am Fam Physician, 2007. 75(1): p. 56-63.
4. Vincent, R.G., et al., *The changing histopathology of lung cancer: a review of 1682 cases.* Cancer, 1977. 39(4): p. 1647-55.
5. Tamura, T., et al., *Specific organ metastases and survival in metastatic non-small-cell lung cancer.* Mol Clin Oncol, 2015. 3(1): p. 217-221.
6. Zhou, Q., et al., *[Screening and establishment of human lung cancer cell lines with organ-specific metastasis potential].* Zhongguo Fei Ai Za Zhi, 2014. 17(3): p. 175-82.
7. Pankova, D., et al., *RASSF1A controls tissue stiffness and cancer stem-like cells in lung adenocarcinoma.* Embo j, 2019. 38(13): p. e100532.
8. Heng, W.S., R. Gosens, and F.A.E. Kruyt, *Lung cancer stem cells: origin, features, maintenance mechanisms and therapeutic targeting.* Biochem Pharmacol, 2019. 160: p. 121-133.
9. Liu, A., H. Saman, and P. Pusalkar, *Unilateral temporal haemianopia in a patient with non-small cell lung cancer: intraocular metastasis despite chemotherapy.* BMJ Case Rep, 2011. 2011.
10. Lin, L., J. Sun, and J. Wang, *Lung cancer and intraocular metastasis in gestation: Clinical experiences of a rare case.* Thorac Cancer, 2020. 11(9): p. 2723-2726.
11. Nelson, C.C., B.S. Hertzberg, and G.K. Klintworth, *A histopathologic study of 716 unselected eyes in patients with cancer at the time of death.* Am J Ophthalmol, 1983. 95(6): p. 788-93.
12. Kreusel, K.M., et al., *Choroidal metastasis in disseminated lung cancer: frequency and risk factors.* Am J Ophthalmol, 2002. 134(3): p. 445-7.
13. Su, H.T., Y.M. Chen, and R.P. Perng, *Symptomatic ocular metastases in lung cancer.* Respirology, 2008. 13(2): p. 303-305.
14. Liu, W.C., et al., *Using Machine Learning Methods to Predict Bone Metastases in Breast Infiltrating Ductal Carcinoma Patients.* Front Public Health, 2022. 10: p. 922510.
15. Lynch, C.M., et al., *Prediction of lung cancer patient survival via supervised machine learning classification techniques.* Int J Med Inform, 2017. 108: p. 1-8.
16. Lu, M.T., et al., *Deep Learning Using Chest Radiographs to Identify High-Risk Smokers for Lung Cancer Screening Computed Tomography: Development and Validation of a Prediction Model.* Ann Intern Med, 2020. 173(9): p. 704-713.
17. Nageswaran, S., et al., *Lung Cancer Classification and Prediction Using Machine Learning and Image Processing.* Biomed Res Int, 2022. 2022: p. 1755460.
18. Luengo, J., et al., *Addressing data complexity for imbalanced data sets: analysis of SMOTE-based oversampling and evolutionary undersampling.* Soft Computing, 2011. 15(10): p. 1909-1936.
19. Gramegna, A. and P. Giudici, *SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk.* Front Artif Intell, 2021. 4: p. 752558.
20. Lundberg, S.M., et al., *From Local Explanations to Global Understanding with Explainable AI for Trees.* Nat Mach Intell, 2020. 2(1): p. 56-67.
21. Schmidt, N., C. Neumann-Haefelin, and R. Thimme, *Cellular immune responses to hepatocellular carcinoma: lessons for immunotherapy.* Dig Dis, 2012. 30(5): p. 483-91.
22. Meng, W., et al., *The immunosuppression role of alpha-fetoprotein in human hepatocellular carcinoma.* Discov Med, 2016. 21(118): p. 489-94.
23. Zhu, M., et al., *Hepatitis B Virus X Protein Driven Alpha Fetoprotein Expression to Promote Malignant Behaviors of Normal Liver Cells and Hepatoma Cells.* J Cancer, 2016. 7(8): p. 935-46.
24. Gao, F., et al., *Predictive value of tumor markers in patients with recurrent hepatocellular carcinoma in different vascular invasion pattern.* Hepatobiliary Pancreat Dis Int, 2016. 15(4): p. 371-7.

25. Tang, J., et al., *Clinical Significance of CYFRA21-1, AFP, CA-153, CEA, and CA-199 in the Diagnosis of Lung Cancer Ocular Metastasis in Hypertension Population.* Front Cardiovasc Med, 2021. 8: p. 670594.

26. Čamdžić, N., et al., *Serum total prostate-specific antigen (tPSA): correlation with diagnosis and grading of prostate cancer in core needle biopsy.* Med Glas (Zenica), 2021. 18(1): p. 122-127.

27. Ge, Q.M., et al., *Ocular Metastasis in Elderly Lung Cancer Patients: Potential Risk Factors of CA-125, CA-153 and TPSA.* Cancer Manag Res, 2020. 12: p. 1801-1808.

28. Bottoni, P. and R. Scatena, *The Role of CA 125 as Tumor Marker: Biochemical and Clinical Aspects.* Adv Exp Med Biol, 2015. 867: p. 229-44.

29. Wang, C.F., et al., *The Combination of CA125 and NSE Is Useful for Predicting Liver Metastasis of Lung Cancer.* Dis Markers, 2020. 2020: p. 8850873.

30. Bind, M.K., et al., *Serum CA 19-9 and CA 125 as a diagnostic marker in carcinoma of gallbladder.* Indian J Pathol Microbiol, 2021. 64(1): p. 65-68.

31. Duffy, M.J., et al., *Biomarkers in Breast Cancer: Where Are We and Where Are We Going?* Adv Clin Chem, 2015. 71: p. 1-23.

32. Li, B., et al., *CA-125, CA-153, and CYFRA21-1 as clinical indicators in male lung cancer with ocular metastasis.* J Cancer, 2020. 11(10): p. 2730-2736.

33. Jian, L., et al., *Electrochemiluminescence immunosensor for cytokeratin fragment antigen 21-1 detection using electrochemically mediated atom transfer radical polymerization.* Mikrochim Acta, 2021. 188(4): p. 115.

34. Huang, Y.L., et al., *Diagnostic accuracy of cytokeratin-19 fragment (CYFRA 21-1) for bladder cancer: a systematic review and meta-analysis.* Tumour Biol, 2015. 36(5): p. 3137-45.

35. Wu, H.H., et al., *Serum cytokeratin-19 fragment (Cyfra 21-1) is a prognostic indicator for epithelial ovarian cancer.* Taiwan J Obstet Gynecol, 2014. 53(1): p. 30-4.

36. Shirasu, H., et al., *CYFRA 21-1 predicts the efficacy of nivolumab in patients with advanced lung adenocarcinoma.* Tumour Biol, 2018. 40(2): p. 1010428318760420.

37. Muley, T., et al., *Potential for the blood-based biomarkers cytokeratin 19 fragment (CYFRA 21-1) and human epididymal protein 4 (HE4) to detect recurrence during monitoring after surgical resection of adenocarcinoma of the lung.* Lung Cancer, 2019. 130: p. 194-200.

38. Lin, Q., et al., *Diagnostic value of CA-153 and CYFRA 21-1 in predicting intraocular metastasis in patients with metastatic lung cancer.* Cancer Med, 2020. 9(4): p. 1279-1286.

39. Minamibata, A., et al., *Age and Smoking Status Affect Serum Cytokeratin 19 Fragment Levels in Individuals Without Cancer.* In Vivo, 2022. 36(5): p. 2297-2302.

40. Wang, Y., et al., *Short-term load forecasting of industrial customers based on SVMD and XGBoost.* International Journal of Electrical Power & Energy Systems, 2021. 129: p. 106830.